OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1038/543007a

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Denk, F. (2017). Don't let useful data go to waste. *NATURE*, *543*(7643), 7. https://doi.org/10.1038/543007a

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Don’t let useful data go to waste

A boon of open science is going unutilized — researchers must seek out others’ deposited findings, urges Franziska Denk

One of my favourite parts of being a scientist is attending conferences – preferably in sunny climes away from rainy London. But recently, my joy at being in Sicily or Japan has been tempered by the same frustration, time and again: too few of my fellow neuroscientists seem to be consulting my data – or others’. What is going on?

Terabytes of biological information have been accumulated via next generation sequencing (NGS) and other high-throughput methods. Many scientists have used some of those resources, designed to be widely re-purposed, such as ENCODE. But many more laboratories in neuroscience or other subdisciplines of cell and molecular biology generate their own NGS datasets for a particular purpose. For instance, I study microglial cells which are thought to be important for the maintenance of chronic pain. Immunologists may look at these same microglia in their work on multiple sclerosis. We all generate cell-type specific transcriptional results that, among other information, provide a full profile of which genes are being expressed in naïve microglia.

These studies are often not part of a concerted effort, but for them to be published, we must still deposit them into a public repository for anyone to download. In the case of NGS data, it is usually the so-called Gene Expression Omnibus (GEO) website run by NCBI, which will issue a unique identifying number known as GEO accession. This means that every biologist can now find out what a microglia should look like from a molecular perspective. The same is true for a great many other kinds of cells, for instance neurons or all cell types found in human blood (http://blueprint-data.bsc.es/release_2016-08/#!/).

This is useful information. Knowing whether your favourite gene is being actively used in a certain cell provides important clues about how to proceed with your research. That’s why funders and journals have worked so hard over the past decade to ensure that data-generating researchers like me make our results public instead of guarding them jealously like Gollum and his Precious.

Hence my surprise when sitting through long discussions at conferences about whether gene X is found in microglia, when I can quickly open a few Excel files on my laptop and see that it is absent from this cell type in relevant RNA sequencing datasets. Similarly many papers I read or review make claims about which proteins are expressed or not which don’t match publicly available transcriptional results.

Of course people should not take others’ data as gospel. Sequencing data can be wrong. There can be systemic technical artefacts. And known biases are associated with certain approaches, for example single cell datasets can miss more than half the transcripts in individual samples. But simply taking no notice of deposited data is akin to ignoring several independently published replication experiments. If your results don’t agree, you should, at the very least, discuss the discrepancy, and propose a biologically valid reason for it.

Why are many bench biologists ignoring this new wealth of cell-type specific expression data? After all, any scientist working in cell or molecular biology is likely to require such information frequently. My hunch is that there are two reasons: researchers underestimate how many of these data have been published over the past few years; and they are wary of the information derived from them. Because you need bioinformatics knowledge to generate and analyse NGS data, people assume that they also need such expertise to locate and interpret the results. Not so. In the last five years, improvements in NGS technologies and stricter data deposition guidelines mean papers are commonly accompanied by simple Excel files that can be downloaded in minutes and require minimal knowledge to browse. They can either be directly obtained from the Supplement of a relevant paper or can be found under the "GEO Datasets" tab on NCBI using search terms: like pubmed, but for spreadsheets.

It is often difficult to share big data in science. NGS is fairly unique, in that it is easy to standardize, display and judge from the outside. This is not the case for many other kinds of scientific output. For instance resources for data sharing in brain imaging (Eklund et al., 2016) or engineering (Wallis et al., 2013) are less developed. Obstacles include the high cost of storage – though valiant efforts have been made, for instance in 3D neuronal anatomy (Ascoli et al., 2017).

What a shame, then, to let this information go to waste. More researchers need to be made aware that they can profit from a vast library of material — from transcript levels to regulatory and epigenetic information, such as histone modifications. To that end, I've made a little step-by-step guide and a video on how to access public NGS data (https://www.franziskadenk.com/resources/). You will find links to purpose-built websites, such as Blueprint, which allow you to simply enter the name of your gene – akin to a genome browser, only that it will return cell-type specific transcriptional data. And you will be able to find instructions on how to download and interpret Excel data files from GEO.

Journals could also help disseminate this kind of information, by requiring scientists to state categorically that they have checked their own claims on gene expression against several publicly available sequencing results. And reviewers could verify these statements: spending 15 min searching for a few spreadsheets on GEO is not much different to spending 15 min on pubmed to confirm other types of statements on prior literature.

The data are out there – and many more are to come. All we have to do is access them.